

UNCLASSIFIED

Assessment of 2020 Census Data Quality Processes

JASON Study co-Leads

Sallie Keller, Distinguished Professor of Biocomplexity, University of Virginia

Dan Meiron, Aeronautics and Applied and Computational Mathematics, Caltech

February 23, 2021



UNCLASSIFIED

Why come to JASON Scientific Advisory Group?

2015

Respondent Validation for Non-ID Processing
in the 2020 Decennial Census

2016

*Alternative Futures for the Conduct
of the 2030 Census*

2019

Formal Privacy Methods for the 2020 Census

2020

Secure Computation for Business Data

Contact: Gordon Long — glong@mitre.org

JSR-20-2E

November 2020

DISTRIBUTION A. Approved for public release. Distribution is unlimited.

JASON
The MITRE Corporation
7715 Colshire Drive
McLean, Virginia 22102-7508
(703) 903-6907




Founded in 1959, JASON is a select group of scientists and engineers who advise the U.S. government on scientific and technical matters related to national security.

JASON operates autonomously, is self-governed, and is not obligated to please.

Many members of U.S. National Academies of Sciences, Engineering, and Medicine.

JASON has history of advising Census Bureau.

Putting the 2020 Census in context



Census Bureau brings together more than a decade of planning to conduct a complete and accurate enumeration of each person in the United States.

As in other decades, the Census Bureau developed a detailed plan for the enumeration and subsequent data processing, one that included flexibility for contingencies.

Unprecedented disruptions raise special concerns over the possible erosion of data quality of the 2020 Census.

The question before JASON

“What should the Census Bureau take action upon to strengthen production and release of metrics to inform on 2020 Census data quality?”

January 4th to 7th of 2021 JASON was briefed on census operations in the field, and the post-data collection processing that was currently underway for the delivery of resident population counts to support Congressional Apportionment and Redistricting Data Products.



First Enumeration for 2020: Toksook Bay, Alaska

JASON report completed February 8, 2021.

Every decennial census is different



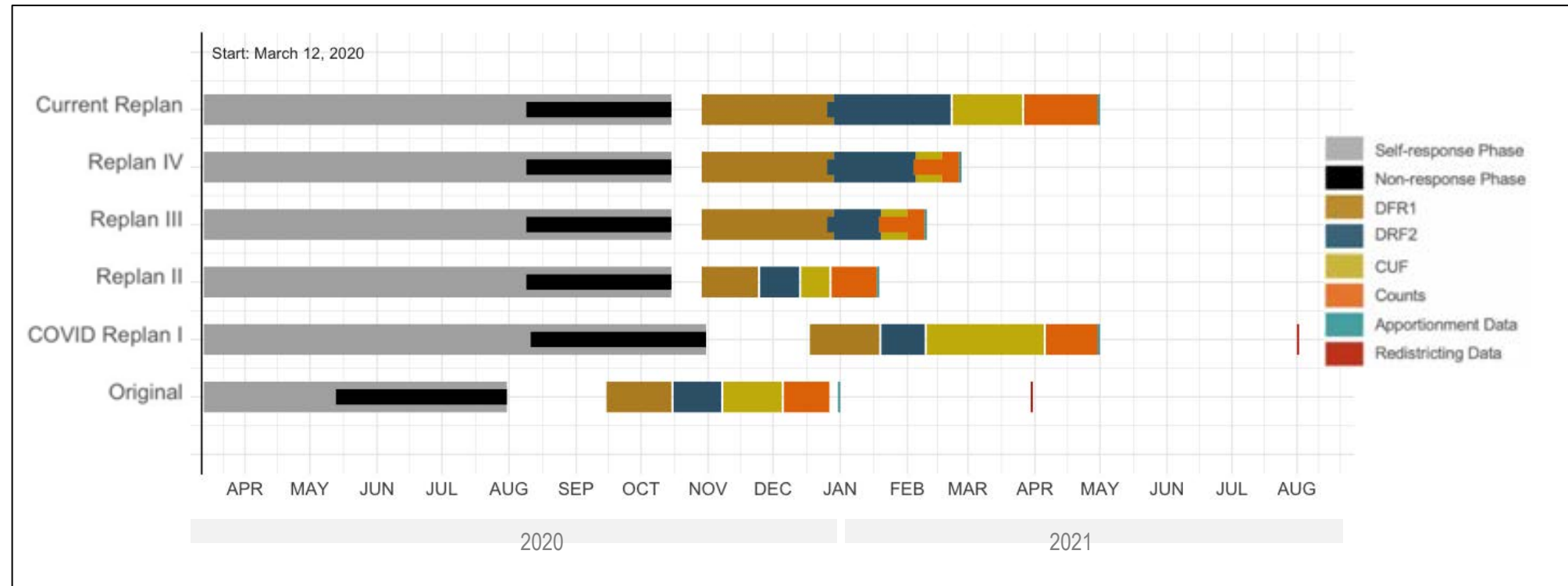
U.S. Census Bureau

It is neither possible nor desirable to repeat the prior process.

A reasonable standard is whether the *plan* for the decennial census would produce results that *improve* on the standards set in previous decades.

It is important to understand—and for the **Census Bureau to communicate to the public**—that the accepted range of results over previous decades allows for considerable imperfections, as long as these do not knowingly embody *a priori* biases against individual states or statutorily defined classes of individuals.

A range of externalities led to revisions of the timeline for delivery of census results



Decennial Response File 1 (DRF1) – processes executed to determine the universe of housing units and general quarters, identification of unique persons, standardization of responses, and a quality assessment of the non-response follow-up (NRFU).

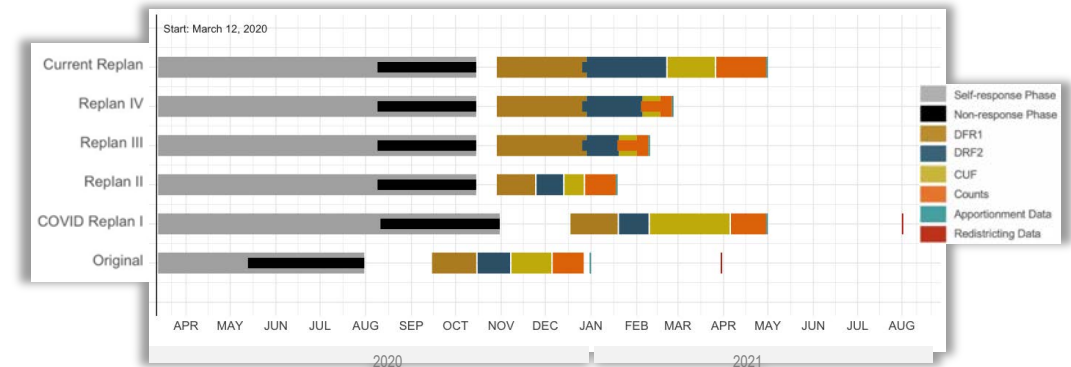
Decennial Response File 2 (DRF2) – deduplication of responses.

Census Unedited File (CUF) – integration of administrative records and determination of the final population count for each address.

Counts – state level population counts created.

Census field operations and post-data collection processing are deliberate activities that take time

The potential consequences of these schedule changes lie at the heart of questions arising around data quality.



Recommendation: In accordance with Sec. 4 of President Biden’s Executive Order (“Ensuring a Lawful and Accurate Enumeration and Apportionment Pursuant to the Decennial Census”, January 20, 2021), the Secretary of Commerce should order a return to a timeline that allows adequate time for complete, accurate, and transparent processing of the data.

Measuring the quality of the process

The Census Bureau would like to know if the choices they made around schedule and the associated changes to its operations have led to a biased or differential count of the population.

Would some other well-informed group have made similar decisions?

This is a counterfactual that cannot be tested.

Quality metrics can be developed to examine whether the enumeration process was reasonable and fair:

- Metrics that capture spatial or demographic differences
- Metrics that provide direct comparisons to various baselines
- Tracking metrics that capture implications of operational shifts over time

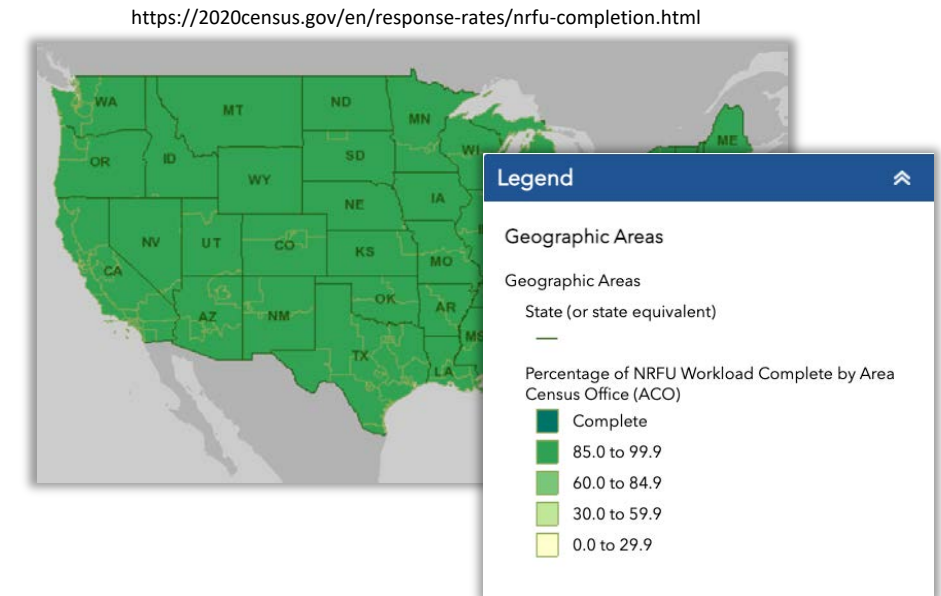
Intra-comparisons of quality metrics

Census Bureau has proposed several data quality metrics and **JASON agrees these are useful. But more can be done.**

Making a distinction between *process measures* that quantify the steps taken during data collection and *outcome measures* that quantify the collected data may be helpful.

- **Process measures** include number of enumerators per area, average number of NRFU visits, and average time conducting interviews
- **Outcome measures** include % of completed enumerations, % of self-responses, % of enumerations by proxy

This distinction allows correlating process and outcome measures as one way to assess data quality.



Intra-comparisons will be useful, but must avoid unintentional disclosure of private data

Formal Privacy Methods for the 2020 Census

Contact: Gordon Long — glong@mitre.org

April 2020

JSR-19-2F

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

JASON
The MITRE Corporation
7515 Colshire Drive
McLean, Virginia 22102-7508
(703) 983-6997



For any public release of information about data quality, disclosure analysis is required to determine if a result can be released or if it is necessary to add privacy noise and count the release against the overall privacy loss budget.

Recommendation: To gain confidence around potential differential count of the population the Census Bureau should make use of its data science resources and summarize the assessments of data quality across various geographies and for relevant demographic groups. Any public release from these analyses should be limited to summary statistics that do not require consumption of privacy loss budget. The expenditure of privacy loss budget should be prioritized to optimize accuracy of future Census Bureau data product releases.

Inter-comparisons of data quality metrics

Inter-comparison of data quality measures involves direct comparison of the data against some (at least partially) independent control set of data such as previous censuses or other demographic surveys.

The Census Bureau plans such assessments through the Demographic Analysis, comparisons with 2010 Census, and through the Post Enumeration Survey.

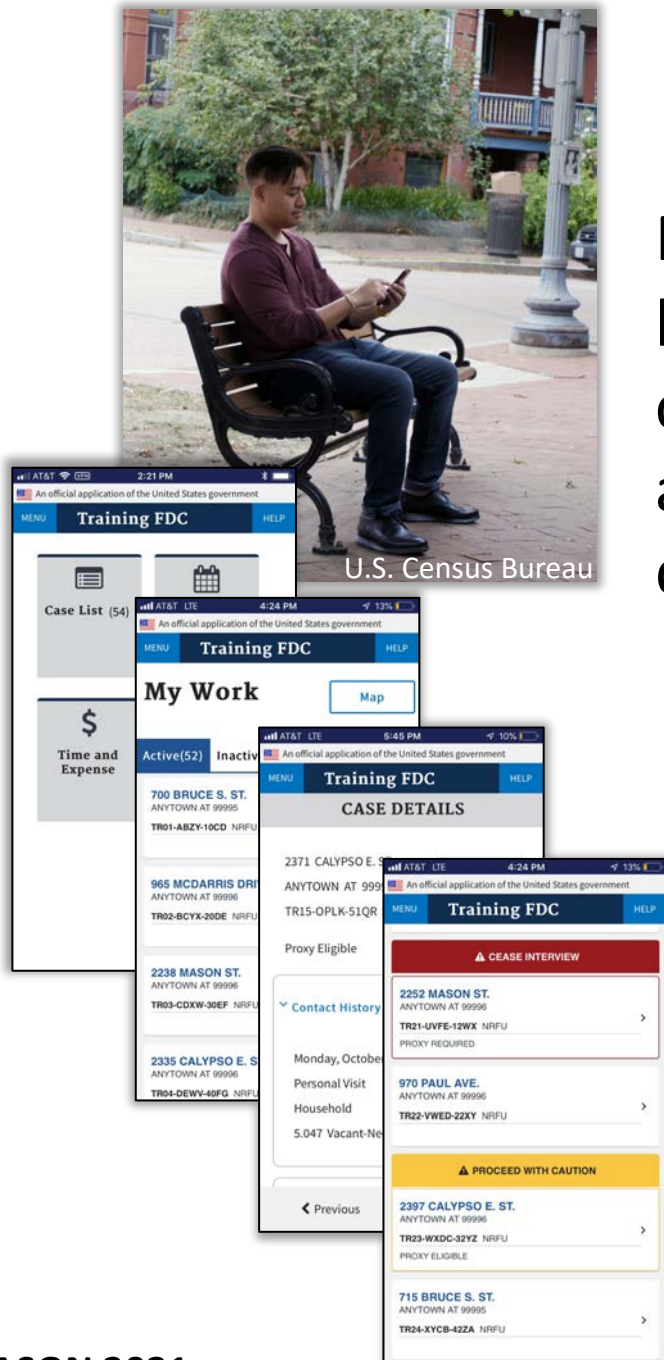
Recommendation: JASON recommends the Census Bureau evaluate the direct comparisons between the 2020 Census and the 2019 Census Test response rates during this current data quality assessment phase with a focus on effects of the response to COVID-19.



Innovations for the 2020 Census

Planned innovations for 2020 helped mitigate some enumeration challenges presented by COVID-19 and the displacement of people due to environmental conditions.

- Increased use of administrative records and third-party data in both address canvassing, NRFU interview scheduling, and enumerations.
- Internet response mode added.
- Reengineering of field operations management allowed for rapid assessment of activities and redistribution of resources.



Time series reflecting field operations are feasible

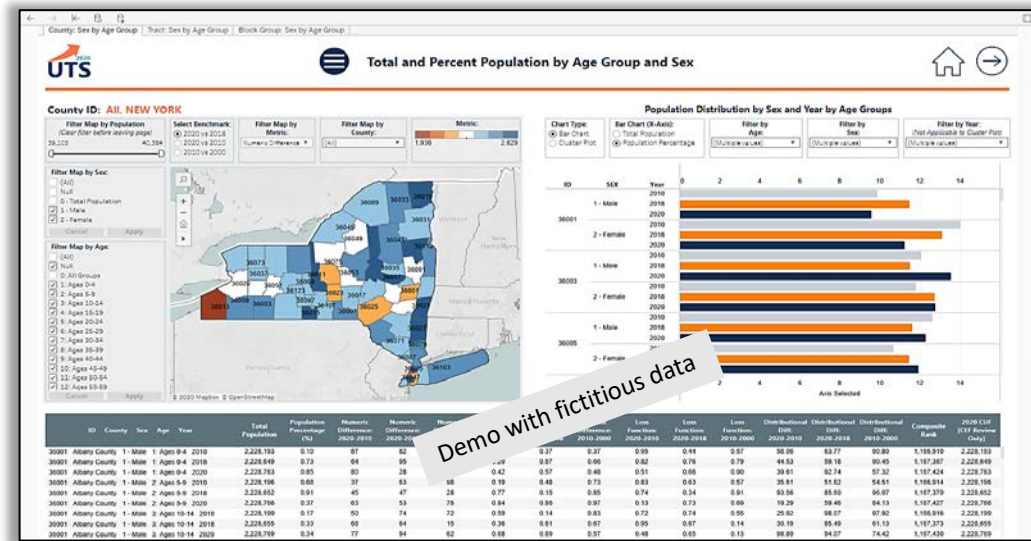
The Census Bureau developed an impressive information technology infrastructure via development of software and hardware that made it possible to monitor the data collection process and to adapt deployment of resources in response to issues.

- Survey Operational Control System (SOCS)
- Field Operational Control System (FOCS)
- Field Data Collection App (FDCA)



Recommendation: Census Bureau should collect and analyze the data from its control system tools. This data should then be summarized to develop indicators of enumeration progress as a function of time that can be compared to progress measures in previous censuses. Depictions of these indicators that highlight responses by Census Bureau to the pandemic and other events should be developed for communication to relevant stakeholders.

New tool: Census Review Analysis and Visualization Application (CRAVA)



The Census Bureau always performs in-depth reviews of the quality of the data.

For 2020 the CRAVA tool was developed to assess the quality of the data at various stages of post-collection back-end processing **to support a team of internal subject matter experts** in the end-to-end review of the data.

Recommendation: Census Bureau should make use of the reports generated by the CRAVA tool to show how data quality issues were identified and addressed over time as the various Census Bureau data products (DRF1, DRF2, CUF, etc.) were produced. These issues and their resolution could be summarized graphically, using a set of time-dependent quality metrics obtained from retrospective analysis of the enumeration data and processes.

~~Anomalies~~ are expected issues

Census Bureau performs several checks for data completeness and consistency and detects potential inconsistencies that the Census Bureau terms anomalies.

Such inconsistencies are not surprising or indicative of serious problems but are an expected and normal part of the data quality-control processes.

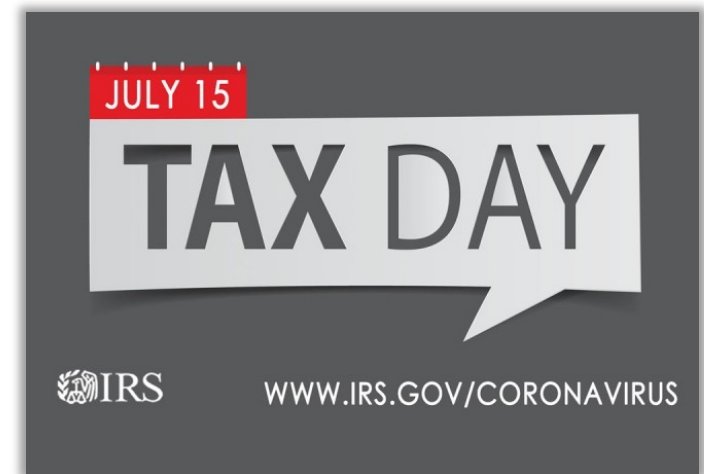
Recommendation: The Census Bureau should change the terminology around detected issues from specialized terms that can be misconstrued as negative judgments when they are actually useful quality measures. The Census Bureau should consider avoiding using the term “anomaly”, which may raise inappropriate alarm to non-experts, and communicate openly about the measures used to check for data consistency, the specific issues uncovered, and steps taken to address them. If possible, comparison should be made with the number and type of issues identified during the 2010 Census.

Administrative records use was also impacted

As planned, the 2020 Census expected to make more use of administrative records than in the 2010 decennial.

The planned uses of administrative records were modified for the COVID-19 disruptions and the IRS tax filing delay.

Recommendation: Looking forward to the 2030 Census JASON recommends that the Census Bureau explore more opportunities to use administrative records and third-party data, while maintaining transparency and careful quality controls on any use of such data.





Communications could be improved

Communication has focused on the importance of being counted, advertising and use of Census data products, and Census research.

These efforts are laudable.

However, the Census Bureau **falls short** in their communication efforts surrounding the enumeration processes and changes made to these processes, including those resulting in improvements!

Recommendation: JASON recommends the Census Bureau develop tiered messaging ranging from videos of a few minutes to a longer presentation (~1 hour) that provides increasing levels of quantification of the myriad efforts undertaken to ensure an accurate count in the 2020 Census.

Conclusions and next steps



The 2020 Census faced unprecedented challenges.

Innovations for 2020 Census allowed Census Bureau to pivot activities quickly in response to unplanned scheduling.

Census Bureau's new internal systems for capturing information could be leveraged to enhance their quality assessments.

Census Bureau has capabilities for excellent communications, and these should be engaged in better communicating the high quality of the enumeration processes.